DOCUMENT RESUME

ED 467 372 TM 034 298

AUTHOR	Sotaridona, Leonardo S.; Meijer, Rob R.
TITLE	Statistical Properties of the K-Index for Detecting Answer Copying. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-01-06
PUB DATE	2001-00-00
NOTE	32p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Cheating; Sample Size; Simulation; *Statistical Analysis
IDENTIFIERS	*K Index; *Type I Errors

ABSTRACT

This study investigated statistical properties of the K-index (Holland, 1996) that can be used to detect copying behavior on tests. A simulation study was conducted to investigate the applicability of the K-index for small, medium, and large datasets. In addition, the Type I error rate and the detection rate of this index were compared with the copying index of J. Wollack (1997). Several approximations were used to calculate the K-index. Results show that all approximations were able to hold the Type I error rates below the nominal level. Results further show that using the copying index results in higher decision rates than the K-indices for small and medium sample sizes (100 and 500 simulees). (Contains 6 figures and 14 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 467 372

TM

Statistical Properties of the K-index For Detecting Answer Copying

Research Report 01-06

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Leonardo S. Sotaridona
Rob R. Meijer

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM034298

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

BEST COPY AVAILABLE

Statistical Properties of the K-index for Detecting Answer Copying

Leonardo S. Sotaridona
Rob R. Meijer

Abstract

We investigated the statistical properties of the K-index (Holland, 1996) that can be used to detect copying behavior on a test. A simulation study was conducted to investigate the applicability of the K-index for small, medium, and large datasets. Furthermore, the Type I error rate and the detection rate of this index were compared with the copying index, ω (Wollack, 1997). Several approximations were used to calculate the K-index. Results showed that all approximations were able to hold the Type I error rates below the nominal level. Results further showed that using ω resulted in higher detection rates than the K-indices for small and medium sample sizes (100 and 500 simulees).

Key Words: IRT, nominal response model, copying indices, cheating

The variety of methods to cheat on educational tests seems to be only restricted to one's imagination. In his book on cheating on tests, Cizek (1999, Chap. 3) gives an overview of several cheating methods. Among the methods discussed are using forbidden materials, circumventing the testing process, or even using microrecorders.

In the present study, we will be concerned with a form of cheating that has received some attention in the recent literature, namely, answer copying. In this type of cheating, one examinee copies the answers from another examinee. This copying may take place from an examinee who is sitting in the neighborhood of the copier, although answer copying may also take place using all kinds of codes for transmitting answers and a code for doing so, for example, clicking of pens, tapping of the foot, and the like. Thus the examinees do not have to be in the physical neighborhood of each other. Because answer copying may invalidate an examinee's test score, it is necessary to prevent those practices by using well-instructed proctors and construct the seating arrangements so that there is ample room between the examinees. However, if a proctor observes some irregularities, statistical methods may be used to obtain additional evidence of answer copying.

Several methods have been proposed that all are based on determining the probability that the observed score patterns of two examinees under suspicion are similar. If this probability is high, this may indicate that one examinee copied the answers from another examinee. These chance methods can be classified into two types (Cizek, 1999, pp. 138-139). One type of method compares an observed pattern of responses to a known theoretical distribution (e.g., Frary, Tideman, & Watts, 1977; Wollack, 1997). In the second type of method, the probability of an observed pattern is compared with a distribution of values derived from independent pairs of students who took the same test. An example of such a statistic is the K-index (Holland, 1996).

In this paper we will investigate the statistical properties and the detection rate of the K-index which thus far is only described in a paper by Holland (1996) and applied on a few empirical datasets from Educational Testing Service (ETS). As Cizek (1999) noted, no comparative studies of the performance of this index are known, so it is unknown whether it performs better, worse, or the same as the other available indices. In this paper

we will investigate the statistical properties of the K-index and, in particular, the small sample properties of this index. Furthermore, we will compare the detection rate of this index with the index, ω, proposed by Wollack (1997). The major difference between the indices is that the K-index does not assume any test model, whereas ω is based on item response theory modeling (e.g., van der Linden & Hambleton, 1997).

This study is organized as follows. First, we will discuss the rationale behind the K-index and discuss several methods proposed by Holland (1996) to calculate this index. Second, we will discuss some existing practical problems when this index is applied in practice and we will propose two new methods to calculate this index. Third, we will conduct a simulation study to investigate the statistical properties of this index and finally, we will conduct a simulation study in which we compare the Type I error rate and detection rate of the K-index with the ω statistic.

The K-index

The K-index is a statistic that can be used to assess the degree of unusual agreement between the incorrect answers on a multiple-choice test of two examinees; one referred to as the *source* (s) and the other as the *copier* (c). The copier is suspected of copying answers from the source. Note that the K-index only takes the *in*correct answers of the examinees into account. For a rationale behind this strategy, see Holland (1996).

Notation

The following notation will be used throughout the text. Let

j ($j = 1, \ldots, J$) denotes examinees,

i ($i = 1, \ldots, I$) denotes items,

v ($v = 1, \ldots, V$) denotes the item response categories,

s denotes an examinee identified as the source,

c denotes an examinee suspected of copying answers from s,

w_j denotes the number of "wrong" answers of examinee j,

M with realization m denotes the number of matching wrong answers between examinee j and s,

$r = 1, \ldots, c', \ldots, R$ denotes subgroups of examinees, where each group has a distinct number of wrong answers and c' is the group where examinee c belongs

$j' = 1, \ldots, n_r$ denotes an examinee in subgroup r, where each subgroup has at least one examinee and $\sum_{r=1}^{R} n_r = J - 1$,

$\mathbf{M}_r = (M_{r1}, \ldots, M_{rj'}, \ldots, M_{rn_r})$ denotes a vector of matching wrong answers in a particular subgroup r

$\mathbf{M}_{c'} = (M_{c'1}, \ldots, M_{c'n_{c'}})$ denotes a vector of the number of matching wrong answers of $n_{c'}$ examinees in subgroup c' where subgroup c' consists of the examinees with the same number-incorrect score as the copier,

and let $Q_r = \frac{w_r}{I}$ denote the proportion of wrong answers of subgroup r where I is the total number of items in the test.

K-index Based on the Empirical Distribution

The K-index can be determined using empirical data of J persons answering I items. To calculate the K-index based on the empirical data, we first determine the group of examinees with the same number-incorrect score as the copier (subgroup c') and then for each of these examinees in subgroup c' we determine the number of items that match the incorrect answers of the source. This is the vector $\mathbf{M}_{c'}$ and the distribution of $\mathbf{M}_{c'}$ comprises the empirical agreement distribution. For examinee c, we specifically denote $m_{c'c}$ as the number of matching wrong answers between c and s. The random variable M_{rj} is denoted as M if it is not necessary to identify the group membership of j. The K-index is defined as the proportion of examinees having the *same* number-incorrect score as c whose number of matching incorrect item scores with s is at least as large as $m_{c'c}$.

For $j' = 1, \ldots, n_{c'}$, let $I_{c'j'}$ denote an indicator variable, coded as 1 for $m_{c'j'} \geq m_{c'c}$, and 0 otherwise, then K is defined as

$$K = \frac{\sum_{j'=1}^{n_{c'}} I_{c'j'}}{n_{c'}}. \quad (1)$$

The idea is that when K is very small there is statistical evidence that examinee c copied from examinee s.

Note that, in general, the number of matching incorrect scores depends on the ability level of s and c. The number of matching incorrect answers is necessarily small when either s or c, or both have many correct scores (high ability), whereas it is large when both examinees have many wrong answers (low ability). In order to minimize the dependency of M on the ability level of the population of examinees, the K-index is computed conditional on the number of incorrect scores of the suspected copier. As a consequence, the number of examinees involved in the actual computation of the K-index (subgroup r) becomes very small. We emphasize this because the number of examinees in a subgroup r influences the accuracy of the value of the K-index. When the sample size is small ($J = 100$) one alternative is to use a theoretical approximation to the empirical agreement distribution.

K-index Based on Theoretical Approximations

To use the K-index, one has to specify first the Type I error (α) which is defined as the probability of misclassifying an examinee as a copier. Ideally, we would like to have a statistic for which the nominal and empirical Type I error rates are similar. Note that in this type of statistical application, the main concern is to have a statistic that is not liberal–a statistic for which the empirical Type I error rate is at most as large as the nominal Type I error rate–because the consequence of misclassifying an honest examinee as a copier can be very serious at the individual level.

Seaman et al. (1991; see also, Wollack, 1997) argued that copying indices that fail to hold the nominal Type I error rate should be considered unacceptable. On the other hand, the copying index should not be overly conservative; otherwise, the power of the copying index to detect true examinee copiers will be very low.

In general, a disadvantage of using the discrete empirical distribution in small samples is that the random variable M can only take a small number of values. As a result, it is often not possible to obtain a prespecified Type I error of say .05 (Agresti, 1996, p. 43).

Holland (1996) noted that the distribution of M can be approximated by the binomial distribution, that is: $M \overset{approx.}{\sim} B(w_s, p)$ where w_s, the number of wrong answers of the source is known, but p is unknown. Holland (1996) suggested two ways of approximating p. In the first approach, p is computed such that the binomial distribution and the empirical distribution of M have the same means. Let $\overline{m}_{c'}$ denote the mean of the empirical agreement distribution which equals

$$\overline{m}_{c'} = \frac{\sum_{j'=1}^{n_{c'}} m_{c'j'}}{n_{c'}}. \tag{2}$$

Then, an estimate of p denoted as $p^*_{c'}$ is defined as

$$p^*_{c'} = \frac{\overline{m}_{c'}}{w_s}. \tag{3}$$

Let K^* denote the K-index based on $p^*_{c'}$, then K^* is given by

$$K^* = P(M \geq m_{c'c}) = \sum_{g=m_{c'c}}^{w_s} \binom{w_s}{g} (p^*_{c'})^g (1 - p^*_{c'})^{w_s - g}. \tag{4}$$

Holland (1996) showed using large empirical datasets that the binomial distribution using p^*_c yielded a "conservative" estimate of the empirical agreement distribution. That is, the K-index based on the binomial approximation is often stochastically higher than the K-index based on the empirical distribution (Agresti, 1990, p. 9).

To calculate $p^*_{c'}$, the response pattern of examinees in the subgroup c' must be available. Furthermore, the value of $p^*_{c'}$ is affected by the sample size– the smaller the sample size, the less reliable is the estimate of $p^*_{c'}$. Holland (1996) suggested to

approximate $p^*_{c'}$ through linear regression by utilizing the proportion of wrong answers (Q_r) of each examinee in each number incorrect score subgroup $r = 1, \ldots, R$. Using large datasets from ETS, Holland (1996) showed empirically that p^*_r, where p^*_r is defined analogously as in equation (3), is linearly related to Q_r. Let $\widehat{p}_r$ be the estimate of the binomial probability p^*_r using Q_r. The expression for $\widehat{p}_r$ is given as a piece-wise linear function with a and b as the intercept and slope parameters, respectively:

$$\widehat{p}_r = \begin{cases} a + bQ_r & \text{if } 0 < Q_r \leq 0.3 \\ [a + .3b] + .4b[Q_r - .3] & \text{if } 0.3 < Q_r \leq 1 \end{cases} \tag{5}$$

Note that a and b have to be specified in order to estimate $\widehat{p}_r$ in equation (5). Holland (1996) used $a = 0.085$ and different values for b depending on the particular test that was used. However, from his study it is unclear how these values were obtained. Besides, they may vary across different tests.

In the present study, we will propose $\widehat{p}^*_1$ and $\widehat{p}^*_2$ as estimates of p^*_r based on linear and quadratic regression approach. Based on these estimates of p^*, two versions of K-index, $\overline{K}_1$ and $\overline{K}_2$ are defined as

$$\overline{K}_1 = P(M \geq m_{c'c}) = \sum_{g=m_{c'c}}^{w_s} \binom{w_s}{g} (\widehat{p}^*_1)^g (1 - \widehat{p}^*_1)^{w_s - g} \tag{6}$$

and

$$\overline{K}_2 = P(M \geq m_{c'c}) = \sum_{g=m_{c'c}}^{w_s} \binom{w_s}{g} (\widehat{p}^*_2)^g (1 - \widehat{p}^*_2)^{w_s - g}. \tag{7}$$

Note that only those examinees belonging to subgroup c' are used to estimate p by $p^*_{c'}$. On the other hand, $\widehat{p}^*_1$ and $\widehat{p}^*_2$ use relevant information from R subgroups. Therefore, $\widehat{p}^*_1$ and $\widehat{p}^*_2$ are expected to provide better estimates of p than $p^*_{c'}$.

The main aim of this study is to explore the usefulness of the K-index and its approximations given in equations (4), (6), and (7) under varying testing conditions.

First, we will investigate if the linear relationship between p_r^* and Q_r found by Holland (1996) also applies for relatively small datasets. Second, we will investigate the fit of the binomial distribution using p_c^*, $\widehat{p}_1^*$, and $\widehat{p}_2^*$ as an approximation to the distribution of M. Finally, we will determine the empirical Type I error rates and detection rates of the K-index and the ω statistic (Wollack, 1997). Because we will use ω to evaluate the performance of the K-index, we will introduce this statistic first.

The ω statistic

Wollack (1997) proposed the ω copying index that is formulated in the context of the nominal response model (NRM, Bock, 1972). To determine ω, the NRM is used to estimate the probability that an examinee responds to one of the item response categories $v[= 1, \ldots, h, \ldots, V]$. Under the NRM, the probability of examinee j with ability level θ_j responding to option h of item i with intercept and slope parameters ζ_{ih} and λ_{ih} is given as

$$P_{ih}(\theta_j) = \frac{\exp(\zeta_{ih} + \lambda_{ih}\theta_j)}{\sum_{v=1}^{V} \exp(\zeta_{iv} + \lambda_{iv}\theta_j)}. \tag{8}$$

Let h_{cs} be the number of identically answered items of s and c, let $E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi})$ be the expected value of h_{cs} conditional on the ability level of the copier (θ_c), the item response vector of the source ($\mathbf{U}_s$), and the item parameters ($\boldsymbol{\xi}$). Furthermore, let $\sigma_{h_{cs}}$ be the standard deviation of h_{cs}. Then ω is given by

$$\omega = \frac{h_{cs} - E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi})}{\sigma_{h_{cs}}}, \tag{9}$$

where

$$E(h_{cs}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi}) = \sum_{j=1}^{J} P(u_{jc} = u_{js}|\theta_c, \mathbf{U}_s, \boldsymbol{\xi}).$$

Using the NRM, the probabilities of c selecting the responses of s can be determined. For any pair of examinees s and c, the distribution of ω approaches the standard normal (Wollack, 1997) as the number of test items becomes infinitely large. Thus, the ω values can be evaluated for statistical significance using the standard normal distribution.

The ω statistic is very similar to the g_2 index proposed by Frary et al. (1977). The main difference is in the way the expected value of h_{cs} is computed; ω uses the nominal response model conditional on θ_c, $\mathbf{U}_s$, and $\boldsymbol{\xi}$, whereas g_2 uses item distractors and difficulties from classical test theory and the ratio of the copier's number-correct score to the mean number-correct score for all examinees.

Wollack (1997) compared the empirical Type I error rates and the power of ω and g_2. The results showed that ω performed better than g_2 in detecting answer copying, under the conditions simulated. In particular, g_2 failed to maintain the nominal Type I error rate which he found was too liberal in all circumstances. Therefore, in this study, the empirical Type I error and detection rates of the K-index were compared with ω.

Although both the K-index and ω make use of item response similarities, ω compares the responses of the copier to the entire response vector of the source, whereas in the K-index, the incorrect responses of the copier are compared with the incorrect responses of the source. Wollack (1996, p.13) pointed out that the power of a statistic that does not take into account the information from correctly answered items is likely to be reduced due to a reduction in the number of operational items used. Besides, examinees that are most likely to be caught are those who miss several items. He added that "it is often not worthwhile to pursue a cheating claim if the alleged copier received a low score"; an argument against a copying index that disregards correctly answered items such as the K-index.

The ω statistic is based on IRT modeling, in particular the nominal response model. First, it is reasonable to assume that the fit of the model to the data is important for the ω statistic to perform well. Second, if the suspected examinee copied a considerable number of items from the source, the ability level of the copier will be overestimated which consequently affects the value of ω. Finally, the estimation of the item parameters used

in the NRM requires large number of examinees (Wollack, 1997); a requirement which may restrict the usefulness of this index in cases where large datasets are not available, although Wollack (1998) showed that estimating the item parameters on sample size as small as 100 for 40 and 80 items test did not result in an increase in Type I error or a significant loss in power.

The K-index on the other hand, does not assume any IRT model and is therefore easier to apply in practice. However, a drawback of this index is that the number of examinee in each score group based on the number-incorrect scores should be large enough to obtain a reliable estimate of the binomial p. For example, when simulating 10 times a test consisting of 40 items and drawing θ from the standard normal distribution for 30 simulees, the number of score groups ranges from 19 through 22 with score groups with only 1 simulee ranging from 12 through 15 (60-74%) and other score groups consisting of only 2 or 3 simulees. Thus, p is very unreliably estimated for these samples.

Method

Data Generation

The NRM was used to generate item scores on multiple-choice tests with five options. Test lengths were 40 and 80 items and the number of simulees in the sample were 100, 500, and 2000. These numbers were chosen to reflect small, medium, and large sample sizes. To be able to compare the results in this study with the results obtained by Wollack (1997), the same item parameters were chosen as in his study which were based on empirical data of a mathematics college placement test. Similarly, the ability parameter, θ_j, was drawn from $N(0,1)$. Given the item and ability parameters, $P_{ih}(\theta_j)$ was computed for all i, h and j, using equation (8).

Items with five answer categories were considered. The observed response of examinee j to item i was obtained by drawing a sample from the set $v = \{1, \ldots, 5\}$, where each element of v has a probability of being drawn equal to $P_{i1}(\theta_j)$, $P_{i2}(\theta_j), \ldots, P_{i5}(\theta_j)$ respectively. In the NRM, the category with the largest algebraic value for λ has

a monotonically increasing response function. As in other studies (e.g., Thissen & Steinberg, 1997), this category was chosen as the keyed alternative.

Simulation of Copying

To simulate copying, s and c were identified based on their ability percentile rank. Because in practice we are mainly interested in obtaining additional statistical evidence of answer copying for examinees that raise their scores by copying answers from an examinee with higher ability, we choose c such that the ability percentile rank of c is lower than that of s. This was also done to reflect the fact that the source is often a person with higher ability level than the copier (Holland, 1996). Simulees were first ordered according to θ. Then, in each dataset, the source was selected as the simulee at the 90th or 60th percentile rank. In each dataset, 5% copiers were selected randomly from the simulees with θ level below the θ level of the source.

Similar to Wollack (1997), copying was simulated by first randomly selecting an item and then altering the response of c to match the responses of s. This was done as follows. First n% (e.g., 10%, 20%, 30%, 40%) of the items were randomly selected and then the item scores of c on these items were changed to match the item scores of s. For both 40-item and the 80-item tests, 10%, 20%, 30%, and 40% of the item scores were changed corresponding to 4, 8, 12, and 16 items in the 40-items test and 8, 16, 24, and 32 items in the 80-items test. The four factors – sample size (3 levels), number of items (2 levels), ability level of the source (2 factors), and percentage of items copied (4 levels) – were completely crossed to simulate 48 testing conditions. A program in S-plus (S-PLUS 2000, MathSoft Inc.) was written by the authors that performed the required simulation and necessary routine calculations.

The data used in this study share the following similar features with the data used by Wollack (1997): [1] the copier copied from a more able source; [2] the number of copiers in each dataset and the percentage copied were the same, and [3] the same item parameters and distributional assumption were made for the θ parameters.

A difference with Wollack (1997) is that we did not use a seating chart to identify the $s - c$ pair. We assumed that there is a suspicion that c copied the answers from s.

The K-index and the ω statistic were then used to check the probability that copying has occurred for a particular $s - c$ pair of examinees. So we did not use the statistics as a screening device. Wollack (1996) pointed out that in situations where there is only one source, ω has the highest power.

Data Analysis

Relationship Between p_r^* and Q_r

Recall that Q_r $(r = 1, \ldots, R)$ denote the proportion of wrong answers in each number-incorrect score group. For each score group r, we computed the binomial probability p_r^* using equation (3) with $\overline{m}_{c'}$ replaced by $\overline{m}_r$ which is the mean of the empirical agreement distribution for subgroup r. To explore the relationship between Q_r and p_r^*, we first created scatterplots for p_r^* and Q_r. The information derived from visual inspection of these scatterplot suggested the kind of regression models to be fitted. On the basis of the results discussed below and on the empirical results obtained by Holland (1996), two standard linear regression models were proposed: *(a)* $\widehat{p}_1^* = \beta_0 + \beta_1 Q_r + \varepsilon_r$ and *(b)* $\widehat{p}_2^* = \beta_0 + \beta_1 Q_r + \beta_2 Q_r^2 + \varepsilon_r$, where β_0 and β_1 are the slope and intercept parameters respectively, β_2 is a regression parameter that indicates direction and amount of curvature, and ε_r is an error term which is assumed to have a normal distribution with mean 0 and constant variance σ^2. The fit of the two models was determined using the coefficient of multiple determination (R^2) and the magnitude of the residual standard error (see Neter et al., 1996). R^2 measures the proportionate reduction of total variation in p_r^* associated with the use of Q_r. The model with the largest R^2 and the smallest RSE was preferred.

Type I Error and Detection Rates

For a given α, a simulee was identified as a copier when the value of the K-index was less than or equal to α. For the ω statistic, a simulee was identified as a copier when the value of ω was above the one-tailed critical value corresponding to the upper α of the standard normal curve. In this study, assuming suspicion of a specific simulee copying from a specific source, the ω statistic was tested for significance without adjustment for

α level. α =.0001, .0005, .001, .0025, .005, .01 were used. These values were also used in Wollack (1997).

To investigate the empirical Type I error rate, we simulated tests of 40 and 80 items for 100, 500, and 2000 persons and we computed the number of times a truly noncopier was incorrectly identified as a copier. We used 100 replications. Similarly, the detection rate was investigated by taking the proportion of replications where the true copier c was detected.

Results

Relationship Between p_τ^* and Q_τ

Scatter plots of p_τ^* and Q_τ were investigated for different sample sizes and number of items. Results are shown in Figure 1. For sample size $J = 100$ (Figure 1 a-b), the relationship seems to be linear but for sample size $J = 500$ (Figure 1 c-d) p_τ^* initially increases as Q_τ increases then levels off at approximately $Q_\tau = 0.6$, and tends to decrease. For 2000 examinees (Figure 1 e-f) it is clear that the relationship is curvilinear.

Quantitative assessment of the fit of the *linear* and *quadratic* regression models in terms of R^2 and RSE revealed that the model which included the quadratic term had a better fit, that is, a larger R^2 and a smaller RSE. For example, for $J = 500$ and $I = 40$ (Figure 1c), the value of R^2 for the linear fit is 0.6 ($RSE = 0.03$), whereas including Q_τ^2, the value of R^2 increases to 0.66 ($RSE = 0.03$). Similar observations applied for $J = 2000$. Note that despite the relatively small value of R^2 for $J = 100$, the fit of the quadratic model is still better than the linear model. In general, p_τ^* is estimated more accurately when the quadratic term is included.

Empirical and Binomial Agreement Distributions

For a particular choice of the source and the subject, several agreement distributions were constructed for the empirical K-index and for K^*, $\overline{K}_1$, and $\overline{K}_2$ based on the three versions of the binomial distributions ($p_{c'}^*$, $\widehat{p}_1^*$, and $\widehat{p}_2^*$). Results for different sample sizes

were similar so we present in Figure 2 a typical example of these distributions for sample size 500.

In general, the empirical distribution (Figures 2a) tends to have larger upper tail (negatively skewed) whereas the distribution based on $\widehat{p}_2^*$ (Figures 2d) consistently have smaller upper tails. Note that the size of the upper tail of the distribution greatly influences the value of the K-index. As can be seen from equations (1), (4), (6), and (7), the K-index is computed as the sum of the upper tail probability densities. This implies that a distribution with the smallest upper tail yields smallest numerical values of the K-index and thus provides the strongest evidence of answer copying. Since the empirical agreement distribution has a larger upper tail, it is expected that the K-index computed based on this distribution will be large and thus implies low detection rates.

Further, we found that the empirical distribution had the largest upper tail when the number of simulees was smallest, that is for $J = 100$ (graph not presented here). Thus, for $J = 100$, the K-index based on equation (1) is expected to be too conservative.

Type I Error Rate

Figure 3 shows the graphical comparison of the empirical Type I error rates of the K-index and ω, across combinations of examinee sizes and number of items. Type I error rates that are on the identity (boundary) line represents perfect Type I error control, Type I errors above the boundary line are larger than the nominal values and those below it are smaller than the nominal values. $\overline{K}_1$ and $\overline{K}_2$ (denoted in Figure 3 as K1 and K2, respectively) are K-indices based on equations (6) and (7), whereas the K^* is based on equation (4). The Type I error rate of the K-index based on equation (1) was found to be much below the nominal α level and is not presented here.

The K-indices were able to control the Type Error rates below the nominal alpha level in all situations considered. In most cases, ω was also able to control its Type I error below the nominal level, with the exceptions for the 80-item test with 500 and 2000 simulees wherein the Type I error of ω exceeded its nominal level by approximately .005 (see Figures 3d and 3f).

We also investigated the variance of the K-index and ω across replications. The variance of the K-index decreased with increasing percentage of copied answers, sample size, and number of items. The variance of ω decreased with increasing percentage of copying but unlike the K-index, ω seems not sensitive to changes in sample size and number of items. The variance of the K-index was almost equal to ω for longer tests, large number of examinees, and a large percentage of copying. For example, for an 80-item test and 100 examinees, the variance of $\overline{K}_1$, $\overline{K}_2$ and ω for 10% copying are .0955, .0939 and .0704 respectively. As the percentage of copying increases to 40%, the three variances decrease to .003, .003, and .002, respectively.

Detection Rate

The detection rates of K^*, $\overline{K}_1$, $\overline{K}_2$, and ω as a function of α-level for different percentages of copying, sample sizes and test lengths were first investigated for the source fixed at the 90th percentile. The K-index based on equation (1) was not included in the current analysis because its detection rate was extremely low. Figure 4 shows the detection rates for 100 simulees on the 40-item test and Figure 5 for 500 simulees on the 80-item test. The detection rates for the other simulated configurations were similar and are not presented here.

In almost all simulated datasets, ω had the highest detection rate. The difference between the detection rates of ω and the K-indices is relatively large for small sample size and test length but tends to diminish as the sample size and test length increased. For example, the difference in detection rate between ω and $\overline{K}_2$ is 0.15 for $J = 100$, $I = 40$, and 40% copying (see Figure 4a) and it reduce to 0.02 for $J = 500$, $I = 80$, and 40% copying (see Figure 5a). The K-index based on the binomial distribution where p was estimated using linear regression with quadratic term included ($\overline{K}_2$), appeared to be slightly better than $\overline{K}_1$. As expected , K^* had the lowest detection rate.

Further note that the detection rates of the ω and the K-indices increased with the percentage of copied answers. Thus, examinees who copied many items are more likely to be detected than examinees who copied few items.

The probability of detecting a copier who copied 10% of the items is very low–at most .08 for ω and less than .05 for the K-indices (see Figures 4d and 5d)

Increasing the number of simulees had no substantial effect on the detection rates of ω. This is expected since the computation of ω depends only on the response pattern of the source and the copier and not on other examinees. On the other hand, the detection rates of the K-indices increased with the sample size and number of items. For example, for 40% copying the detection rate of $\overline{K}_2$ is 0.69 for $J = 100$ and $I = 40$ (see Figure 4a) and it increased to 0.92 for $J = 500$ and $I = 80$ (see Figure 5a).

To investigate the influence of the proficiency level of the source, we also investigated the detection rates of the indices when the source was at the 60th percentile rank. Results are shown in Figure 6 for 100 simulees and a 40-item test. Comparing Figure 6 with Figure 4 revealed a slight increase in the detection rate of ω, $\overline{K}_1$ and $\overline{K}_2$ for 40% and 30% copying but for 20% and 10% copying, the detection rates were almost the same; the detection rate of K^* substantially increased for 40% copying but not for the other percentages of copying. Comparing the indices within Figure 6 revealed that ω still maintains the highest detection rate followed by $\overline{K}_1$ and $\overline{K}_2$ which are close to each other and then by K^*.

Discussion

In this study we investigated the statistical properties of the K-index and compared its detection rate with the detection rate of the ω statistic. The practical usefulness of these statistics will depend on the application at hand. As was shown in this study, the use of these indices need not be restricted to large-scale testing but can also be applied for small samples consisting of 100 examinees. As others have discussed, these indices can be used to obtain additional evidence for answer copying when a proctor has observed irregular behavior. An alternative is to use these indices for routine monitoring of test responses to prevent copying or for triggering the need to employ such measures. For example, a faculty member can inform privately a pair with a very high index value of its occurrence and suggest that they not sit together on subsequent tests.

Results showed that in general, the binomial success probability, p, is better estimated by a quadratic function than by a linear function of the proportion wrong answers, Q. However, when the dataset is large ($J = 2000$), the relationship between p and Q was nearly linear at the lower end of Q (e.g., $Q < 0.6$). This finding supported the findings by Holland (1996) when he used the linear function to estimate p by Q. In his study, he used ETS data for which the source and the copier generally belonged to the upper end of the ability continuum (e.g., few wrong answers or low value of Q).

When using the K-index for small datasets ($J = 100$), it is not advisable to use the empirical agreement distribution nor its binomial approximation based on equation (4). In terms of distributional shape, the empirical agreement distributions was negatively skewed whereas the binomial distributions–especially the one based on $\widehat{p}_2^*$–exhibited a positively skewed distribution. This resulted in a larger numerical value of the K-index despite the higher percentage of answers copied by the copier.

Results further showed that all approximations of the K-index were able to hold the Type I error rates below the nominal level in all situations simulated. Thus, the K-index has more favorable statistical properties than the g_2 index (Frary et al., 1977) which failed to control the nominal Type I error rates (Wollack, 1996).

Although ω had higher detection rates than $\overline{K}_1$ and $\overline{K}_2$ for simulee sizes 100 and 500, the differences in detection rates are small using 2000 simulees. It is expected that using more that 2000 simulees the detection rates of $\overline{K}_1$ and $\overline{K}_2$ will further improve. We don't recommend to use K^* in practice while $\overline{K}_2$ might be a good alternative if for some reason it is not possible to use ω.

Finally, the random variable M is a non-negative count of matching incorrect answers. For future study, it may be important to investigate the fit of a Poisson distribution as an alternative distribution for the random variable M. Furthermore, the weighted matching correct answers between the source and the copier can be included in the computation of the copying index. The weight may be taken as some function of the probability of correct response. Incorporating the weighted matching correct answers in addition to matching incorrect answers differentiates the K-index from this new index.

Also, several measures can be investigated to minimize the impact of discreteness due to small sample size.

References

Agresti, A. (1990). *Categorical data analysis*. NY: Wiley.

Agresti, A. (1996). *An introduction to categorical data analysis*. NY: Wiley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 46,* 443-459.

Cizek, G. J. (1999). *Cheating on tests: how to do it, detect it, and prevent it.* NJ: Lawrence Erlbaum.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 6,* 152-165.

Holland, P. W. (1996). Assessing unusual agreement between the incorrect answers of two examinees using the K-index: statistical theory and empirical support (*ETS Technical Report No. 96-4*). Princeton, NJ: Educational Testing Service.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th Edition), Mass.:McGraw-Hill

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: some powerful and practicable procedures. *Psychological Bulletin, 110,* 577-586.

S-Plus 2000 *Programmer's Guide and Software*, Data Analysis Products Division, MathSoft, Seatle, WA.

Thissen, D., & Steinberg, L. (1997). A response model for multiple choice items. In: W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). NY: Springer-Verlag.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern item response theory*. NY: Springer-Verlag.

Wollack, J. A. (1996). Detection of answer copying using item response theory (Doctoral dissertation, University of Wisconsin, Madison). *Dissertation Abstracts International, 57/05,* 2015.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21,* 307-320.

Wollack, J.A. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22,* 144-152.

List of Figures

Figure 1. Scatter Plots of p* and Proportion Wrong (Q).

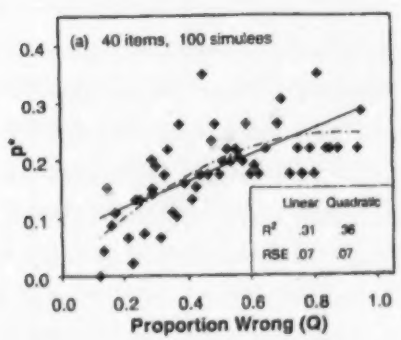
Figure 2. Empirical and Binomial Agreement Distributions.

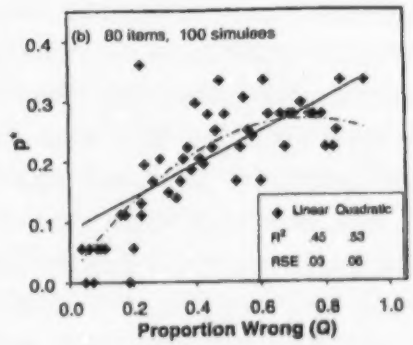
Figure 3. Nominal and Empirical Type I Error Rate as a Function of Simulee Size and Test Length.

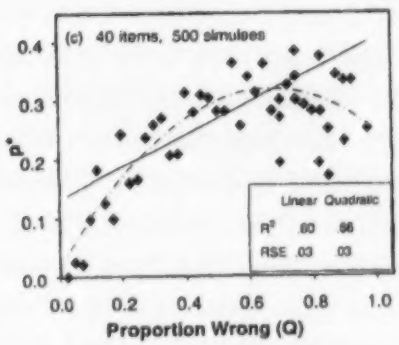
Figure 4. Detection Rate of the K-index and ω, as a Function of Copying Percentage, on 40-item Test, 100 Simulees, and the Source at the 90th Percentile Rank.

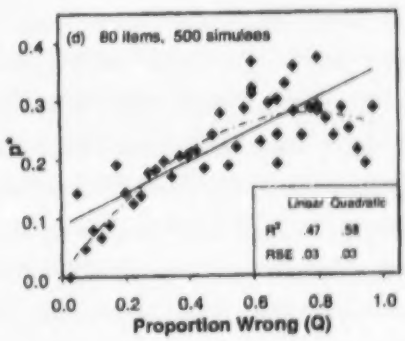
Figure 5. Detection Rate of the K-index and ω, as a Function of Copying Percentage, on 80-item Test, 500 Simulees, and the Source at the 90th Percentile Rank.

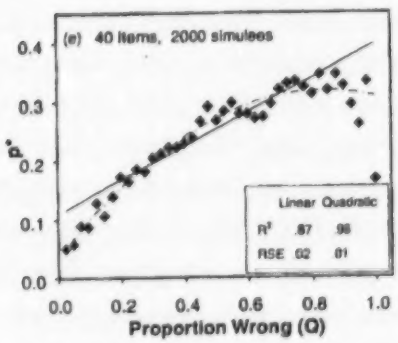
Figure 6. Detection Rate of the K-index and ω, as a Function of Copying Percentage, on 40-item Test, 100 Simulees, and the Source at the 60th Percentile Rank.

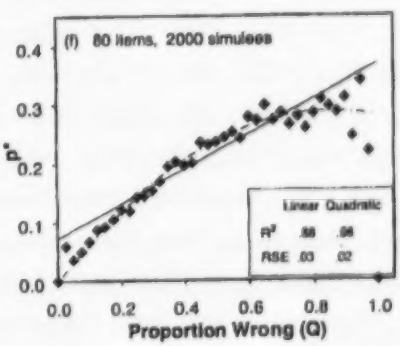




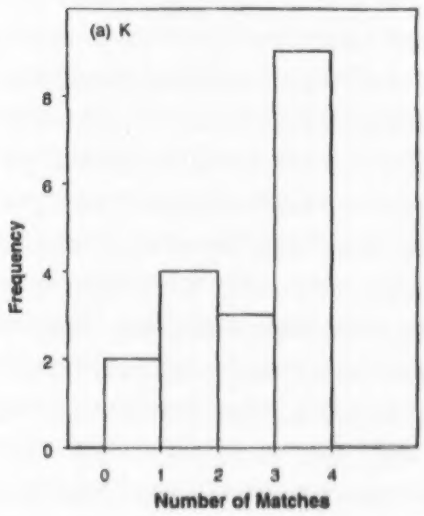


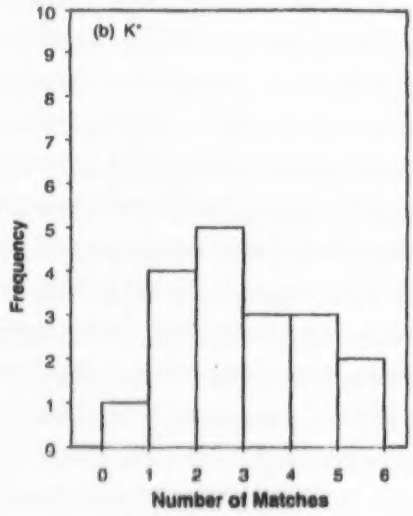


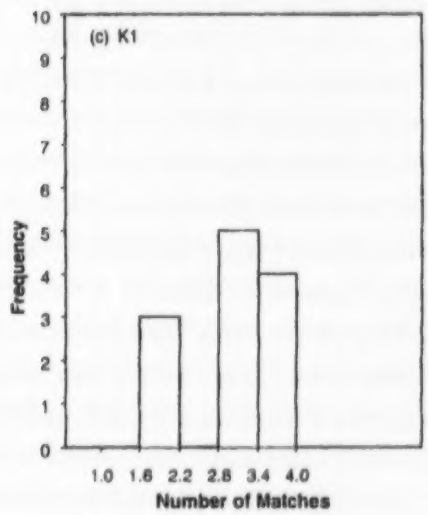


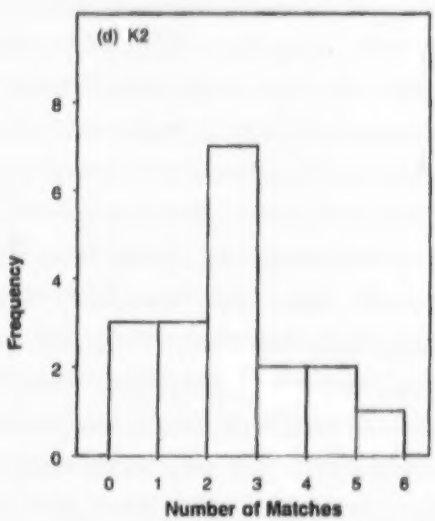


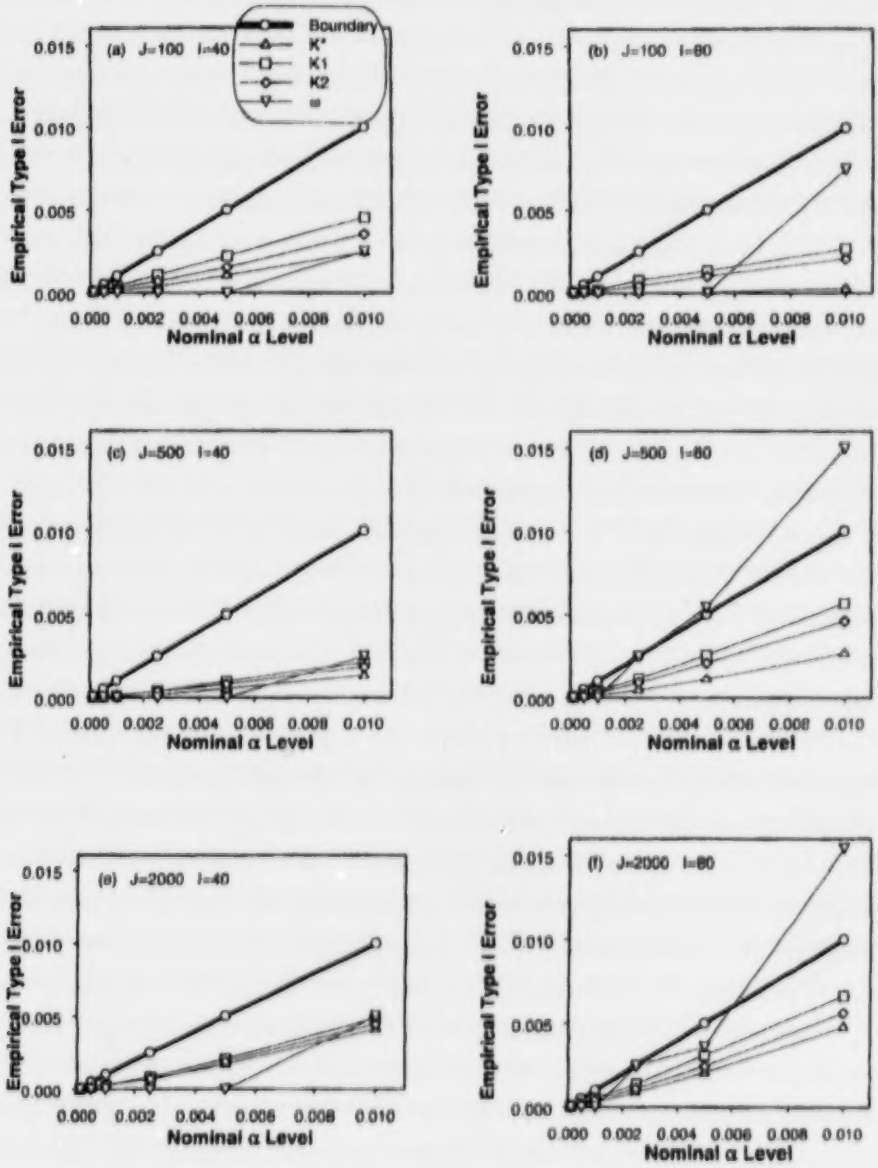
BEST COPY AVAILABLE

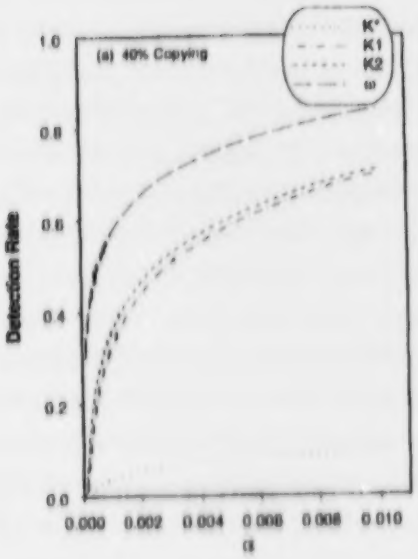

(a) K
Frequency
0
2
4
6
8
0 1 2 3 4
Number of Matches

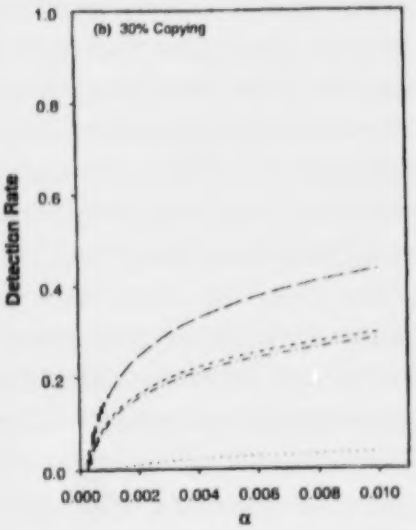

(b) K*
Frequency
0
1
2
3
4
5
6
7
8
9
10
0 1 2 3 4 5 6
Number of Matches

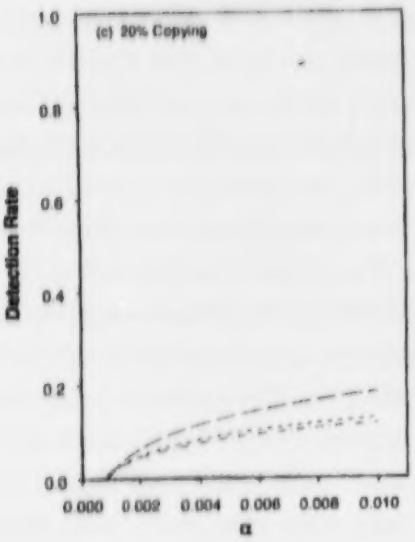

(c) K1
Frequency
0
1
2
3
4
5
6
7
8
9
10
1.0 1.8 2.2 2.8 3.4 4.0
Number of Matches

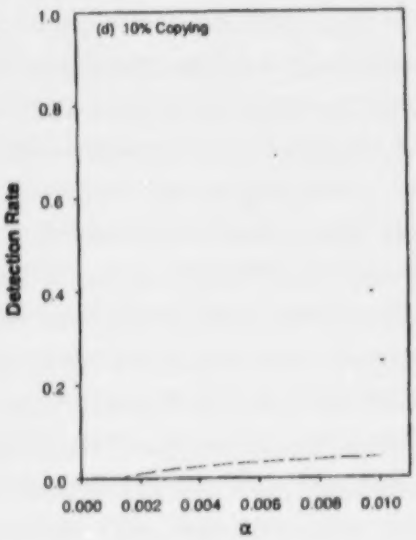

(d) K2
Frequency
0
2
4
6
8
0 1 2 3 4 5 6
Number of Matches

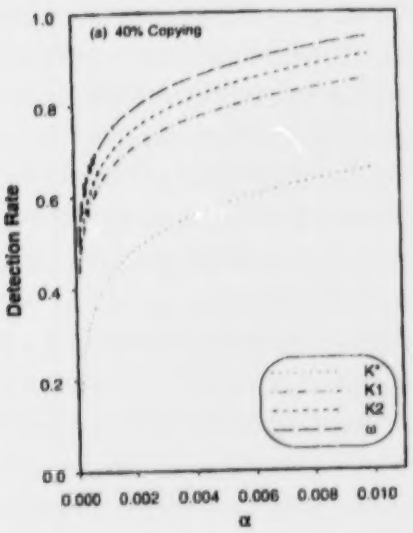

(a) J=100 I=40
(b) J=100 I=80
(c) J=500 I=40
(d) J=500 I=80
(e) J=2000 I=40
(f) J=2000 I=80
Boundary
K*
K1
K2
∞
Empirical Type I Error
Nominal α Level
0.000
0.005
0.010
0.015
0.000 0.002 0.004 0.006 0.008 0.010

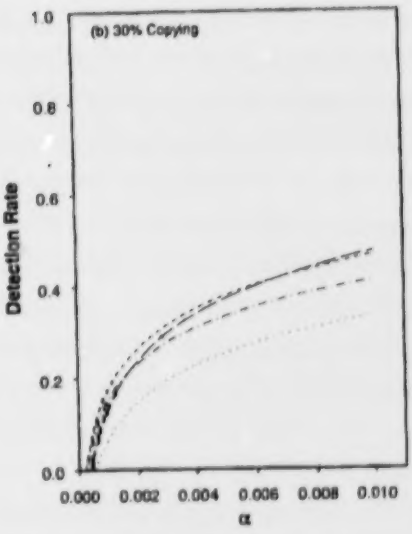

(a) 40% Copying
K*
K1
K2
ω
Detection Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.000 0.002 0.004 0.006 0.008 0.010
α

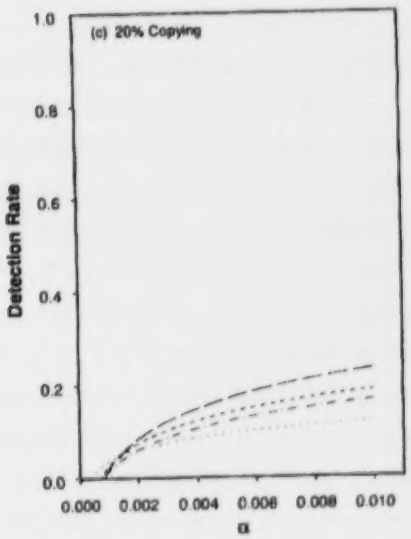

(b) 30% Copying
Detection Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.000 0.002 0.004 0.006 0.008 0.010
α

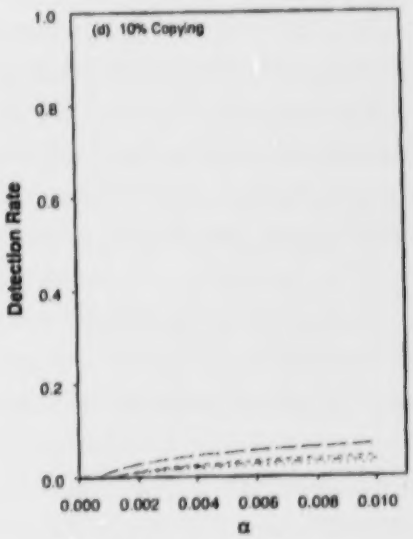

(c) 20% Copying
Detection Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.000 0.002 0.004 0.006 0.008 0.010
α


(d) 10% Copying
Detection Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.000 0.002 0.004 0.006 0.008 0.010
α









BEST COPY AVAILABLE

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-01-06 L.S. Sotaridona & R.R. Meijer, *Statistical Properties of the K-index for Detecting Answer Copying*

RR-01-05 I. Hendrawan, C.A.W. Glas, & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*

RR-01-04 R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*

RR-01-03 R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*

RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 W.J. van der Linden & H. Chang, *Implementing Content Constraints in Alpha-Stratified Adaptive testing Using a Shadow test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

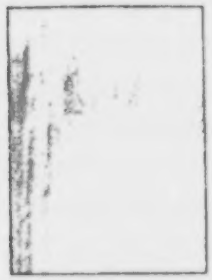
RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.



faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

BEST COPY AVAILABLE

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").